* Measure of dispersion *

Q) Why central tendency is not enough ?

- Consider two diff community in the city which have 5 doctors each, given below is the no. of patients each doctor attends to in a day:

D1	D2	D3	D4	D5
5	2	17	8	18

Community 1

D1	D2	D3	D4	D5
10	9	11	8	12

Community 2

- The avg no. of patients that a doctor attends to in both communities in the same (10), however, all doctors in community 2 seem to be equally busy as opposed to community 1 where a few are very busy and others are not.
- From this we can understand that to completely explain in nature of a data set, measures of central tendency viz. mean, median & mode may be sufficient. so what are the other options?

↳ Answer is we can use Dispersion

- Measure of dispersion are representative of how squeezed or stretched the data is in comparison with the measures of central tendency.
- Measure of dispersion help us identify how our data is spread overall in the data set.
- Measure of dispersion is a non-negative real number which grows as the diversity of data grows.

- Commonly used measures of dispersion are Range, variance, standard deviation, and interquartile range.

① Range :-

- The range is the simplest measure of dispersion that aims at providing an idea of how much the data varies in the data set.
- The range is expressed as.

range = Maximum value in a data set − Minimum value in a dataset.

using community example here as well.

- The range of a no. of patients in community 1 is (18-2), whereas that of community 2 is (12-8)
- The range is used to indicating the spread of the data, in our case, this indicates that the data of community 1 is more dispersed than that of community 2.

② Variance :-

- The range may be inadequate when looking at larger data sets which contain extreme values (ex min. & max.)
- Variance is a better indicator of dispersion. It describes the deviation of data from each other and from the mean.
- variance for a population of size N. can be calculated by using the following formula.

$$\text{variance} = \frac{(x_1-\mu)^2 + (x_2-\mu)^2 + (x_3-\mu)^2 + \dots + (x_n-\mu)^2}{N}$$

$$= \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{N}$$

i) sample variance, for a sample of size n, is

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

ii) Population variance, for a population of size N, is

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

iii) A numerically more stable way (calculator's formula) to calculate sample variance is :

$$s^2 = \frac{1}{n-1} \left(\Sigma(x^2) - \frac{\Sigma(x)^2}{n} \right)$$

Note:

Q) Why for sample variance it is taken as $(1/n-1)$ where as for population it is taken as $(1/N)$

→ It has to do with sample mean, Because sample mean dont have to be actual mean, so the deviation might not be correct, hence reducing the denominator by 1 will increase the sd a little (that also depends on value of n, if n is large it will have less effect and vice versa).

Eg- We are using same community example for variance as well

D1	D2	D3	D4	D5
5	2	17	8	18

Community 1

variance = 51.5

D1	D2	D3	D4	D5
10	9	11	8	12

Community 2

variance = 2.5

- We can observe that variance is much better ~~than~~ indicator of dispersion in comparison with range.

$$\mu = \frac{5+2+17+8+18}{5}$$

$$\mu = 10$$

Community 1

$$\text{variance } (\sigma^2) = \frac{1}{5}\sum_{i=1}^{5}(x_i - \mu^2)$$

$$(\sigma^2) = \frac{1}{5}\left((5-10)^2 + (2-10)^2 + (17-10)^2 + (8-10)^2 + (18-10)^2\right.$$

$$= \frac{(-5)^2 + (-8)^2 + (7)^2 + (-2)^2 + (8)^2}{5}$$

$$\sigma^2 = \frac{25 + 64 + 49 + 4 + 64}{5} = 51.5$$

(3) **Standard Deviation :-**

- The std deviation is the square root of the variance. ~~which~~

$$s = \sqrt{\frac{\Sigma(a - \bar{x})^2}{n-1}}$$

- The dispersion of data can be expressed by either the std deviation or the variance

- ~~For~~ For population std deviation is represented by σ and for sample, the std deviation is represented by s.

Eg

D1	D2	D3	D4	D5
5	2	17	8	18

community 1

variance = 51.5

std deviation σ = $\sqrt{\text{variance}}$

$\sigma = \sqrt{51.5}$

$\sigma = 7.176$

D1	D2	D3	D4	D5
10	9	11	8	12

community 2

variance = 2.5

$\sigma = \sqrt{2.5}$

$\sigma = 1.581$

(3) Z-Score :-

E.g. Doctor 1 graduated from University A in the year 2015 with an aggregate of 78%. whereas Doctor 2 graduated from University B in the year 2015 with an aggregate of 76%. Both these doctors apply for a post at the government hospital in your city.

- The recruiters now need to compare their marks to decide on the candidate to select.
- We have the following info^m with us about the marks of the universities.
  - For marks scored in university A, the mean is 72 and the std deviation is 4.
  - For marks scored in university B, the mean is 68 and the standard deviation is 3.

Q. How can we compare data from these different data sets?

→ Comparision of data from different data sets is done by observing how the data compares against the mean in the data set that it originates from.

This comparision can be performed using z-score, which indicates the distance of the data from the mean in terms of std deviations.

E.g. A z score of 2 indicates that the data is greater than the mean by 2 std deviation, while a score of -1.3 indicates that the

data is lesser than the mean by 1.3 std deviations.

- z-score of a specific value (x) is computed using the formula.

$$z = \frac{(x - \text{mean})}{\text{std deviation}}$$

$$z = \frac{x - \mu}{\sigma}$$

- z-score of doctor 1 is $\frac{(78 - 72)}{4} = 1.5$

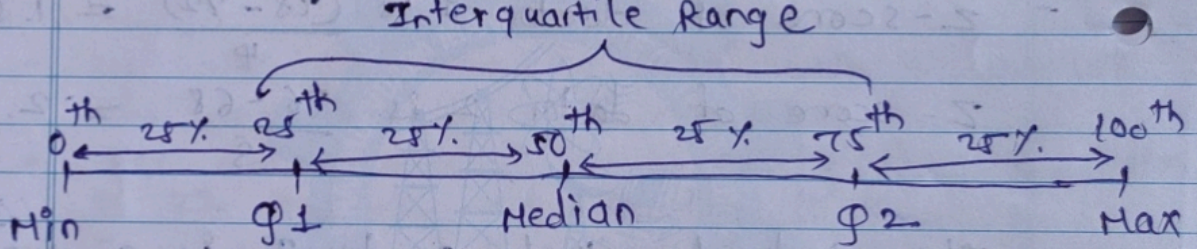
z-score of doctor 2 is $\frac{76 - 68}{3} = 2.33$.

This indicates that Doctor 2 is a more suitable candidate in this scenario

(4) Quartiles :- Percentile

E.g: Given below is the data that represents the average no. of patient in a day for 20 doctors of a specific community.

~~D_1~~	~~D_2~~	~~D_3~~	~~D_4~~	~~D_5~~	~~D_6~~	~~D_7~~	~~D_8~~	~~D_9~~	D_{10}	D_{11}	D_{12}	D_{13}	D_{14}	D_{15}	D_{16}	D_{17}	D_{18}	D_{19}	D_{20}
8	12	4	11	15	10	8	7	4	11	6	6	8	7	4	9	16	6	14	6



The five no. summary for given data is. :

step 1 ✓ sort data in ascending order

4, 4, 4, 6, 6 | 6, 6, 7, 7, 8 | 8, 8, 9, 10, 11 | 11, 12, 14, 15, 16

Min = 4
1st Quartile = 6
Median = 8
2nd Quartile = 11
Max = 16

- The interquartile range (IQR) is the diff b/t Q1 & Q3.

$$IQR = (Q_3 - Q_1) = 11 - 6 = 5$$